On the Robustness of Equilibrium Refinements

ON THE ROBUSTNESS OF EQUILIBRIUM REFINEMENTS *

By

Drew Fudenberg
U.C. Berkeley Department of Economics

and

David M. Kreps
Stanford U. Graduate School of Business

and

David K. Levine
UCLA Department of Economics

UCLA Department of Economics
Working Paper #398

April 1986

* We would like to thank Elon Kohlberg and Jean-Francois Mertens for helpful discussion. We are grateful to the National Science Foundation for financial support.
UCLA Department of Economics, U.C. Berkeley Department of Economics, Stanford University Graduate School of Business, respectively.

1. Introduction

Much effort has been devoted recently to refining the notion of a Nash equilibrium. Beginning with Selten's [1965, 1974] notions of perfection, concepts such as properness (Myerson [1978]), sequentiality (Kreps and Wilson [1982]), and stability (Kohlberg and Mertens [1985]) are often invoked or discussed in the literature. The guiding philosophy is that the analyst knows things about the structure of the game that enable him to reject some of the Nash equilibria as unreasonable. The point of this paper is to indicate that the word *know* in the preceding sentence deserves emphasis. Specifically, the analyst creates a model of the situation that is a simplification and (he hopes) an approximation. Suppose that, in the model, the analyst can reject a particular equilibrium outcome using the various refinements, but for models that are arbitrarily "close" to the one created, this outcome cannot be rejected. Unless faith in the model is absolute, it seems wise to have second thoughts about rejecting this outcome.

To show that this problem is no phantom, we carry out the following basic program in this paper. Fix a space of games and a notion of "closeness" for the games in the space. For each game in the space, a Nash equilibria is *strict* if, for each player, the strategy prescribed is a unique best response (in the normal form) to the other players' strategies. This is as formidable a refinement criterion as we can think of, implying, for example, Kohlberg and Mertens' (1985) *hyperstability*. Now ask: which Nash equilibria of a given game are limit points of strict equilibria for nearby games? We call such equilibria *locally strict*. Following the discussion above, we would hesitate to reject any equilibrium that is locally strict, insofar as the sense of closeness specified initially captures our doubts about the exact specification of the game.

In section 2, we take for the space of games all normal form games over a fixed (finite

player and action) normal form, and we take for "closeness" the Euclidean metric on payoffs. The result is that *every* pure strategy Nash equilibrium is locally strict. (This is not quite true for mixed equilibria, but a weakening of the strictness requirement is given that accomodates every equilibrium, pure or mixed.) This is hardly surprising (the proof is obvious), but then the various refinements of Nash equilibrium are motivated for the most part by the analysis of extensive games. In the rest of the paper, we turn to extensive games with the following philosophy: The analyst is certain (in his model) of the "physical" rules of the game — who moves when, with what information about earlier moves, and so on. That is, roughly, a (physical) extensive form is given. But the analyst may entertain doubts about the players' payoffs and/or their knowledge of the payoffs.

The perturbations that arise from such doubts are considered in the work on reputation and incomplete information. (See Kreps, Milgrom, Roberts and Wilson [1982] and Fudenberg and Maskin [1986].) Our theory differs from this work on reputational effects in one important way. The previous work considers the effect of a fixed perturbation on a *family* of repeated games of varying lengths. The typical result there is that if the horizon is sufficiently long, even outcomes that are not Nash equilibria of the unperturbed game are Nash (and even, see below, locally strict) in the perturbed game. In this paper the game is fixed, and the perturbation is allowed to vary (and made to vanish). Hence only Nash equilibrium outcomes of the original game can be locally strict.

Section 3 concerns the program for cases in which the analyst may be unsure of the players' payoffs himself, but knows that these payoffs are common knowledge among the players. That is, an extensive form is fixed, the space of games is the space of all payoffs for the extensive form, and "closeness" is measured by the Euclidean metric on payoffs. We do not get very clean results here: there do exist locally strict equilibria that are themselves not strict, but the class of locally strict equilibria has no apparent simple characterization.

Sections 4 and 5 are the heart of the paper. Here we imagine that the analyst has no doubts about the physical rules of the game, but is not quite certain of payoffs, and is willing to admit the possibility that the players themselves are not quite certain of each others' payoffs. Put another way, close to a given extensive game are games in which players entertain slight doubts about each others' payoffs, and our analyst is not prepared to reject an equilibrium that cannot be rejected in games that are nearby in this sense.

Section 4 deals with a definition of "closeness" under which every (pure) Nash equilibrium of the original game is locally strict. We begin with a motivating example in subsection 4.1. Technical details and the definition of closeness are given in 4.2. This version of closeness allows each player to be uncertain of his own payoff, and to believe that his opponents may (with small probability) have better information about it than he does. In this setting, an unexpected deviation by one player can signal that all players should change their play. In 4.3 we show that, with this notion of closeness, all pure strategy Nash equilibria are locally strict (and mixed equilibria can be accomodated with a minor weakening of the definition).

In section 5 we consider a second notion of closeness, which requires that each player's additional private information relate only to his own payoffs, and each player's additional information must be independent of the others. With this notion of closeness, an unexpected deviation by a player signals only that his *own* payoffs are different than had been anticipated. With this more restrictive notion of closeness, not all (pure) Nash equilibria are locally strict. However all pure strategy equilibria that are strictly trembling hand perfect in the *normal* form are locally strict, including some that are not subgame perfect.

2. Normal form games and payoff perturbations

Fix a finite player, finite action normal form. (We restrict attention throughout to games with finitely many players, each of whom possesses finitely many strategies.) Let $i = 1, ..., I$ index the players, and let $s_i \in S_i$ index the pure strategies of player i. Denote $\prod_{i=1}^{I} S_i$ by S. Let Γ be the space of games over this normal form: We take $\Gamma = R^{I \times S}$, where for $\gamma \in \Gamma$, $\gamma(i, s)$ is the payoff to player i under strategy s.

The set Γ comes endowed with a natural topology, namely the Euclidean topology. So does Σ_i, the space of mixed strategies for player i, and $\Sigma = \prod_{i=1}^{I} \Sigma_i$. (We will use the term *strategy profile* to refer to elements of Σ.)

A Nash equilibrium $\sigma = (\sigma_1, ..., \sigma_I)$ for a given game γ is called *strict* if, for each player i, σ_i is the player's *unique* best response to the strategies $(\sigma_1, ..., \sigma_{i-1}, \sigma_{i+1}, ..., \sigma_I)$ of the other players. Note that only an equilibrium in pure strategies can be strict, according to this definition. Strict equilibria satisfy all the standard refinements; in particular, a strict equilibrium, taken as a singleton set, is *hyperstable* in the sense of Kohlberg and Mertens

[1985]. [1]

Definition. A strategy profile $\sigma \in \Sigma$ is *locally strict in the normal form* for a game γ if there is a sequence of games γ^n (over the same normal form as γ) and a sequence of strategies $\{\sigma^n\}$ such that (i) $\lim_n \gamma^n = \gamma$, (ii) for each n, σ^n is a strict equilibrium of γ^n, and (iii) $\lim_n \sigma^n = \sigma$.

Proposition 1. A strategy combination $\sigma \in \Sigma$ is locally strict in the normal form for γ if and only if it is a pure strategy Nash equilibrium for γ.

The proof is quite simple. To see that any strategy profile that is locally strict in the Normal form for γ is a Nash equilibrium of γ, use the upper hemi-continuity of the Nash equilibrium correspondence. For the converse, consider the perturbation that adds a small amount to the players' utilities at the outcome prescribed by the strategy in question.

It is easy to deal with mixed strategies, if one permits a small extension to the definition of local strictness. Recall from Kohlberg and Mertens [1985] that for any given normal form game, an *equivalent* normal form game is one in which pure strategies that are convex combinations of other pure strategies are added to or deleted from the original game. Corresponding to every strategy profile in an original game are (possibly many) strategy profiles for a given equivalent game. Consider the following modification of the definition of local strictness in the normal form.

Definition. A strategy profile σ for a normal form game γ is *locally strict in equivalent normal forms* for γ if it corresponds to some strategy profile σ' of an equivalent game γ' that is *locally strict in the normal form* for γ'.

Proposition 2. A strategy profile σ is locally strict in equivalent normal forms for γ if and only if it is a Nash equilibrium of γ.

Again the proof is easy. A simple example will illustrate the method of proof. Consider the game depicted in figure 2.1, and the particular equilibrium in which player 2 randomizes equally between L and R. If figure 2.2 we have an equivalent game γ', with the particular

[1] Conversely, any pure strategy equilibrium that is, as a singleton set, hyperstable, will be strict. Hence as long as we restrict attention to pure strategy equilibria, we could use "hyperstable as a singleton set" instead of "strict".

equilibrium strategy for 2 added as a new strategy for player 2. In this game, this added strategy M corresponds to the mixed strategy of player 2 in the original game; the Nash equilibrium in the original game corresponds to a pure strategy equilibrium in the equivalent game, and we can apply the first proposition.

The spirit of this extended definition is that players may derive a little extra utility from playing a particular randomized strategy. This technique can be applied at any point in the development to follow, to extend results from pure to mixed equilibria. Hence we will, in what follows, discuss only the local strictness of pure strategy profiles, leaving it to the reader to fill in extensions to mixed strategies.

3. Extensive games and payoff perturbations

The result that every pure Nash equilibria of γ is locally strict in the normal form (and, hence, is hyperstable as a singleton set) is hardly surprising, since we allow for *any* sequence of (vanishing) payoff perturbations to the normal form game. Recall that, for most normal form games, *every* Nash equilibrium satisfies the standard refinements; the refinements were created initially to deal with games that arise from a specified extensive form. Hence we might wish to permit, as perturbations to an initial extensive game model, only perturbations that respect the structure of that extensive form.

We can imagine that our outside analyst is concerned with a game over an extensive form E, and he entertains no doubts about the physical rules of the game so specified. But still our analyst may have doubts about the full specification of the situation. He may, for example, be slightly unsure of the payoffs to the players or their probability assessments concerning nature's moves. We will suppose that there is no question about the probability assessments. (The usual accounting tricks make this without loss of generality, if there is no question about the support of those assessments.) Thus we can ask: Fix an extensive form E and probability assessments for nature's moves in E. Let Γ be the space of all payoffs for players in E. For a given $\gamma \in \Gamma$, which strategy profiles are locally strict in the following sense?

Definition. Fixing $\gamma \in \Gamma$, a (pure) strategy profile σ of the game γ is *locally strict in the extensive form* if there exist a sequence of payoffs $\gamma^n \to \gamma$ and (normal form) strict

equilibria σ^n of γ^n such that $\sigma^n \to \sigma$.[2]

We know from Section 2 that, for at least some games, there are equilibria that are locally strict in the normal form but not strict themselves. Viewing a given normal form as a simple extensive form, we see then that there will be profiles that are locally strict in the extensive form but that are not strict. It is easy to show that every strategy profile that is locally strict in the extensive form is a pure strategy Nash equilibrium. But not every pure strategy Nash equilibrium over a given extensive form is locally strict in the extensive form. Consider the extensive game in Figure 3.1. (Our system for diagramming extensive form games is taken from Kreps and Wilson [1982].) It should be easy to see that the Nash equilibrium (L, D) is not locally strict in the extensive form; only (R, U) is.

Indeed, for some extensive games, *no* equilibrium is locally strict in the extensive form. The game depicted in figure 3.2 is an example: While (R, Uu) and (R, Ud) (and, more generally, any randomization between these two strategies) are very nice equilibria, neither is locally strict. In fact, for an equilibrium of an extensive game to be locally strict in the Normal form, it is necessary that every information set is reached with positive probability.[3] This is a very simple manifestation of the type of problem that leads Kohlberg and Mertens to define stable components of equilibria, and we could similarly attempt to define locally strict components. We can, however, obtain cleaner and more provocative results by imagining that our outside analyst is a bit more uncertain as to the game. Roughly, if we weaken the distance measure, so that more games are close to a given game γ, then more equilibria σ will be locally strict.

4. General Elaborations

4.1. An example

To motivate this development, we provide a simple example. Consider first the game in figure 3.1 and the equilibrium (L, D). Imagine an outside analyst whose thought about this situation run as follows:

[2] A better definition would require only that the σ^n be strict in the reduced normal form – see the discussion in section 4.3.

[3] In section 4 we look for strictness in the reduced normal form, which requires that the only information sets with zero prior probability are those at which either (i) the player's choice is irrelevant to everyone's payoff or (ii) the player who moves at the information set is solely responsible for the set having zero prior probability.

"I *know* the particular physical rules of this game. Player 1 must first choose between L and R, and player 2 must choose between U and D if R is chosen. And I'm *fairly* sure that the payoffs are as in figure 3.1. But I'm not completely certain of this. I am willing to admit that there is a small chance that the payoffs are quite different from those shown. Moreover, it might not be common knowledge between the players what those payoffs are. That is, the game may have incomplete information, in that players do not know the precise payoffs, although all the players will have priors that give probability close to one for the payoffs in 3.1."

To model this, we introduce a game that is more elaborate than that in figure 3.1, as follows. Nature moves first, selecting one of several "versions" of the game. Each version is distinguished by having the same extensive form as in 3.1 but may have different payoffs. Nature chooses version having payoffs close to those in 3.1 with probability close to one. Each player has a partition over versions of the game, with each respective player being told in which cell of his partition the true version lies.

One such elaboration is given in figure 4.1. Nature picks one of two versions, the first (having prior probability $1-\epsilon$) with payoffs exactly as in 3.1, and the second (having prior probability ϵ) with very different payoffs. Player 1 knows which version nature picks, while player 2 is not told. Is this *elaboration* of the game in 3.1 very different from the game in (3.1) for small ϵ? Our outside analyst, plagued by the sort of doubts expressed above, might not be willing to rule out the possibility that the players perceive the situation as being that in figure 4.1.

The point of this example is now easy to make. For the game in figure 4.1, the equilibrium (L_1R_2, D) is strict in the normal form. Moreover, it remains strict as we decrease ϵ towards zero. This should not be hard to see: Given that player 2 will play D, player 1's choices at his two information sets are both strictly optimal choices. And, given the strategy of player 1, player 2 is given the move only if nature picks version 2, which renders strictly optimal the choice D by 2.

If we were to measure distance so that, as ϵ goes to zero, this game approaches the game in figure 3.1, then we would conclude (in the spirit of earlier discussion) that (L, D)

in 3.1 is locally strict. This is so even though, in 3.1, (L, D) is subgame imperfect. (Note that a nontrivial step is implicit here: In what sense does (L_1R_2, D) approach (L, D) as ϵ goes to zero? A formal criterion will be suggested shortly.)

With this as a prelude, we now develop a general treatment.

4.2. Elaboration perturbations

We begin with a precise definition concerning when one game is a small perturbation of another. We will not develop a formal topology on the space of games, but instead we give a simple *sufficient* condition for games to be close.

Fix an I-player extensive game of perfect recall, E. This prescribes a game tree T (with nodes denoted by t) which is partitioned into sets of nodes T_i that "belong" to the various players, some of which are initial nodes $w \in W$, and terminal nodes $z \in Z$; an initial assessment ρ over W; information sets $h \in H$, with $H(t)$ denoting the information set containing the (nonterminal) node t; actions $a \in A(h)$ at each information set; and a payoff function $u : I \times Z \to R$ assigning utilities to all players.

The kind of perturbation of E that we have in mind is one in which one of N possible "versions" of the game above is selected by nature at the outset, where each version has the game tree of the game above and, except for initial uncertainty as to nature's choice of version, the same information structure. Versions are distinguished by the players' payoffs. And players are unsure (in a general sense) as to which version prevails. Such perturbations of E will be called *elaborations* of E.

To formalize this, imagine a game $\tilde{E}$ of perfect recall built up as follows. A positive integer N is given, together with a probability distribution μ on $\{1, ..., N\}$. In $\tilde{E}$, the game tree consists of an N fold copy of T, or $T \times \{1, ..., N\}$; we use (t, n) to denote the nth copy of the node t. If player i moves at (non-terminal) node t in E (if $t \in T_i$), then i moves at (t, n) for all n; i.e., $\tilde{T}_i = T_i \times \{1, ..., N\}$. The set of initial nodes consists of the N fold copy of W, with the probability of initial node (w, n) given by $\rho(w)\mu(n)$. The utility to player i at terminal node (z, n) is denoted by $u(i, z, n)$. Finally, information sets are composed as follows. For each player i, a partition P_i of $\{1, 2, ..., N\}$ (with cells denoted by $P_i(n)$) is given, and the information set in $\tilde{E}$ with node $(t, n) \in \tilde{T}_i$ is $H(t) \times P_i(n)$. Actions at information sets are inherited in the obvious fashion.

The information structure bears some scrutiny. Player i's exogenously given information concerning which version n is chosen at the outset is described by the partition P_i. That is, whenever it is player i's turn to move at node (t,n); his knowledge about t is given by $H(t)$ and his knowledge about n is given by $P_i(n)$. Figures 3.1 and 4.1 illustrate the basic construction. The game $\tilde{E}$ in figure 4.1 consists of two versions of E in figure 3.1. In the first (upper) version, the payoffs are identical with those in 4.1. In the second (bottom) version, they are quite different. Player 1 learns at the outset which version is chosen – P_1 is the discrete partition. Player 2 learns nothing – P_2 is the trivial partition.

We consider such elaborations $\tilde{E}$ of a game E as being among the possible perturbations of E. For $\tilde{E}$ to be a "small" perturbation, we take it to be sufficient that the payoffs in $\tilde{E}$ should be, with high probability, approximately those in E. Formally, we pose the following criterion.

Convergence Criterion. For a given game E, let $\{\tilde{E}^k\}$ be a sequence of elaborations of E. To say that the sequence approaches E, it is sufficient that:

(i) there is a uniform bound on the number of versions of the original game in each elaboration $\tilde{E}^k$, and a uniform bound on the absolute values of the u^k;

(ii) for each k, i and z, $\lim_{k\to\infty} u^k(i,z,1) = u(i,z)$; and

(iii) $\lim_{k\to\infty} \mu^k(1) = 1$.

Conditions (ii) and (iii) state that, along the sequence, the probability that nature picks a version in which payoffs are asymptotically as in the original game approaches one. The first part of condition (i) is probably unnecessary to support a notion of closeness, but since we are giving a sufficient condition for convergence, it cannot hurt. The second part of (i) is quite important, however: Even if the probability of other versions is going to zero, if there is no uniform bound on the payoffs of those versions, then we can inflate the payoffs in other versions so that, in expected utility terms, players' payoffs overall are anything we wish them to be. With the uniform bound on payoffs, however, (iii) implies that in ex ante calculations of expected payoffs, only the first version (which gives nearly the payoffs of the original game) will loom large in the limit.

Is this a reasonable sufficient condition for closeness of games in the extensive form? The considerations put into the mouth of our outside analyst at the top of this section would argue that it is, although the reader may already see how broad a class of small perturbations this allows. Insofar as the refinements of Nash equilibrium deal with out-of-equilibrium behavior, a small probability *ex ante* of very different payoffs may loom quite large *ex post*. In what follows, we will see that this is so, and the reader should consider carefully whether the uncertainty of an outside analyst may be so great as this.

4.3. Local strictness under general elaborations

As in the previous two sections, we wish to identify (equilibrium) strategies for a given game E that are locally strict, in this case under the notion of closeness developed above. To do so presents two conceptual problems: First, how does one speak of the convergence of strategies, when strategies in an elaborated game are much richer than strategies in the original game? Since we take the point of view that the outside analyst is only able to view actions in the physical game given by E (and not those actions contingent on which version nature selected), one answer is to look at the distribution induced by strategies on endpoints of the "physical" game. That is, given a strategy $\tilde{\sigma}$ for $\tilde{E}$, look at the distribution induced on Z; given a sequence of strategies on various elaborations of a game, ask for convergence of these distributions. Note that even if all elaborations are trivial, in the sense that each has a single version of T, this gives us convergence of strategies only at information sets that are reached with positive probability. Hence this is weaker than the usual convergence of entire strategies. Because of this, we will take the slightly sharper convergence criterion: we ask for convergence of behavior prescribed in an elaborated game at those information sets $\tilde{h}$ that contain the nodes $(t,1)$.

The second conceptual problem arises in cases in which players have multiple information sets in sequence. Say player i moves at an information set h and then, if he chooses action $a \in A(h)$, he moves again (immediately or after other players move) at information set h'. In any strategy for i that prescribes a move other than a at h, i's choice of action at h' is wholely irrelevant to the payoffs of all the players. Assuming multiple choices at h', i will have multiple pure strategies in the normal form that correspond to a choice other than a at h and that vary in the irrelevant choice at h'. No one of these pure strategies could ever be strictly superior to another, in any game of perfect recall. Following

Kohlberg and Mertens, we identify these multiple strategies, and speak of strictness with respect to the *reduced normal form*, in which all strategies that are convex combinations of others are eliminated and, in particular, these identical strategies are identified.

Definition. An strategy σ for the game E is *locally strict under general elaborations* if there is a sequence of elaborations of E, $\{\tilde{E}^k\}$, that converges to E in the sense of the convergence criterion given above, and a sequence of strict (in the reduced normal form) equilibria $\{\tilde{\sigma}^k\}$ for the respective elaborations, such that the behavior prescribed by the $\tilde{\sigma}^k$ at nodes $(t,1)$ in their respective games converges to behavior at node t prescribed by σ.

Proposition 3. A pure strategy profile σ for the game E is locally strict under general elaborations if and only if it is a (pure) strategy Nash equilibrium in E.

Remark. Mixed behavior strategy equilibria are easily handled as in section 2, by passing to an equivalent extensive form in which the mixed strategy becomes "pure", and perturbing payoffs if the player chooses that particular mixture before passing to the reduced normal form.

Proof. We leave to the reader the job of showing that if σ is locally strict, then it is a Nash equilibrium. For the converse, we fix a (pure strategy) Nash equilibrium σ and construct an elaboration $\tilde{E}^k$ as follows.

For each player i and pure strategy s_i for i, design payoffs for i that make strategy s_i strictly dominant for player i throughout the course of play. To do this, at each terminal node $z \in T$, ask how many times $m_i(z, s_i)$ player i had to deviate from s_i if z is to be reached. (For example, when z lies along a path that passes through no information sets of i, $m_i(z, s_i) = 0$.) To give i the strict incentive to follow s_i at every opportunity, set i's payoffs at $-m_i(z, s_i)$.

Letting $\#S_i$ be the number of pure strategies for player i, in each elaboration there are $\prod_{i=1}^{I}(\#S_i + 1)$ versions of the original game. We label these versions by $(r_1, ..., r_I)$, where each r_i is drawn from the set $S_i \bigcup \{0\}$. In version $(r_1, ..., r_i)$ of elaboration k, the payoffs to player i are as follows:

(i) If $r_i \in S_i$, then player i is assigned the payoffs described above that make r_i a strictly dominant strategy.

(ii) If $r_i = 0$ and, for some $j \neq i$, $r_j \in S_j$, then player i is assigned payoffs that make playing σ_i a strictly dominant strategy.

(iii) If $r_i = 0$ and, moreover, $r_j = 0$ for all j, then player i is assigned utility equal to $u(i,\cdot) - m_i(\cdot,\sigma_i)/k$, where u is the originally specified utility function and m_i is the number-of-deviations function sketched above.

The prior distribution on these many versions is set as follows. Version $(r_1,...,r_I)$ of elaboration k has prior probability $\prod_i \mu_i(r_i)$, where $\mu_i(r_i) = 1/(k \cdot \#S_i)$ if $r_i \in S_i$ and $\mu_i(r_i) = (k-1)/k$ if $r_i = 0$. Finally, for his initial information concerning the version selected, at version $(r_1,....,r_I)$, player i is told the value of r_i.

This construction may seem quite complex, but it has a simple interpretation. Each player i is either "crazy" or "sane", and there are as many different ways for i to be crazy as i has pure strategies. The overall chance that i is crazy in elaboration k is $1/k$, divided equally among the many different forms of craziness. A player is told whether he is crazy or sane and, if crazy, the form that his craziness takes. Players' "types" are selected independently. (The reader should be careful here – payoffs of one player will depend on the type of the other, so payoff perturbations are not independent.) If a player is crazy according to a certain pure strategy, then the player has payoffs that cause him to play that strategy at every available opportunity. If a player is sane and some other player is crazy (of any type), then the sane player wants to follow the fixed Nash equilibrium strategy σ_i at every available opportunity. Finally, if all players are sane, then payoffs are as in the original game, with a small "kick" in favor of the fixed Nash equilibrium strategies σ_i.[4]

It should be evident that this sequence of elaborations converges to the originally given game, according to the criterion we have given. We claim, moreover, that in each elaboration, the following strategies $\tilde{\sigma}_i$ are strict in the reduced normal form: Player i follows σ_i if his initial information is $r_i = 0$, and he follows r_i if his initial information is some $r_i \in S_i$. Once this claim is established, we will have the proposition.

[4] This sort of construction, in which every strategy is played with positive probability, is used by Fudenberg and Maskin (1986) to obtain a robust folk theorem for repeated games with long but finite horizons and small levels of incomplete information.

Suppose that some (pure) strategy $\hat{\sigma}_i$ is one of i's best responses to the strategies $\{\tilde{\sigma}_j\}$ of his opponents (in some elaboration k). We wish to establish that $\hat{\sigma}_i$ is precisely $\tilde{\sigma}_i$ except possibly at information sets that i's own actions preclude. By the assumption of perfect recall, the information sets of i are ordered by precedence. Take any information set h for i that is earliest in terms of precedence among all of i's information sets in which $\hat{\sigma}_i$ is different from $\tilde{\sigma}_i$ and in which i's previous actions (which would be the same under $\hat{\sigma}_i$ and $\tilde{\sigma}_i$ because this is an earliest information set) do not themselves preclude h. If no information sets of i satisfy these conditions, we are done. There are three possibilities to consider:

(i) The information set h belongs to a crazy variety of i. Because i himself doesn't preclude h, and because every strategy profile of his opponents is possible under their strategies in $\tilde{\sigma}$ (conditional on whatever type is i), the information set h has positive prior probability. And at that information set, any action other than that prescribed by $\tilde{\sigma}_i$ does strictly worse than $\tilde{\sigma}_i$ (from then on). Hence there is a strategy for i strictly better than $\hat{\sigma}_i$, which contradicts the assumed optimality of $\hat{\sigma}_i$.

(ii) The information set h belongs to the sane variety of i, and it corresponds to an information set that is hit with positive probability under σ in the original game and equilibrium. Then under $\tilde{\sigma}$ it is hit with positive probability in one of two ways: Either we are in the version in which everyone is sane, in which case following $\tilde{\sigma}_i$ is a strict best response for i (recall the "kicker" in defining utilities for this the all-sane version), or we are in a version in which someone else is crazy and i is sane, in which case following $\tilde{\sigma}_i$ is a strict best response for the rest of the game. In either case, there is a strategy for i strictly better than $\hat{\sigma}_i$, and again we have a contradiction.

(The kicker perturbation to utility in the all-sane version would be unnecessary as long as there is more than one player, since then we would have that following $\tilde{\sigma}_i$ is at least weakly best in the all-sane version, strictly best in every someone-crazy version, and the latter would have strictly positive probability. For one player games, however, the latter has zero probability.)

(iii) The information set h belongs to the sane variety of i, it corresponds to an information set that is off the equilibrium path in the original game under the strategy σ, and player

i does not, by his actions, preclude h. Since there is positive prior probability of every other strategy combination by i's opponents, there is positive prior probability that h will be reached. Moreover, Bayes' rule forces i to conclude that some one or more of his opponents must be crazy at this point, at least insofar as they all follow their parts of $\tilde{\sigma}$. Hence it is strictly better for i to follow $\tilde{\sigma}_i$ at this point and henceforth, and we have the same contradiction as in the previous two steps.

Q.E.D.

5. Independent Elaborations

In terms of our outside analyst story, section 4 says that unless one is *certain* that players will not draw unmodelled inferences about their *own* payoffs from the deviations of their opponents, one should not reject *any* (pure) Nash equilibrium. Suppose, though, that our outside analyst is prepared to assert that the model as written captures all of the information that players have about each other; he entertains (only) the possibility that each player may have unmodelled private information about his own payoffs. In this case the class of small perturbations that the analyst would consider is smaller than in 4.2, and the set of locally strict equilibria that result might be smaller as well.

Consider, for example, the game in figure 3.1 and the Nash equilibrium (L, D). We made this equilibrium locally strict by formulating, in figure 4.1, a game in which player 1 received information that was pertinent to player 2's payoffs. Hence, in this elaboration, player 1, by choosing R, was communicating to 2 that 2's best choice lay with D. (This, of course, supposes that 1 chooses L in the top version of the game.) But if we supposed that player 1 could not be given information about 2's payoffs superior to the information that 2 receives, this would be impossible: 2, moving in the information set that contains the node from the high probability version, would *know* that U is dominant for him.

This is not to say that "independent elaborations" (to be defined formally below) will not render locally strict some equilibria that are not themselves strict. Consider the games in figures 5.1 and 5.2. In figure 5.1, we have an equilibrium (Lu, D) which is subgame imperfect. In figure 5.2 we have an elaboration of this game in which there is some uncertainty as to the payoffs to player 1 *only*. Moreover, player 1 is informed about his payoffs. In this game, $(L_1R_2u_1d_2, D)$ is sequential and a further elaboration (see

below) will make it strict, which renders (Lu, D) locally strict. The intuition is that in the alternative, bottom version of the game, player 1 plays R_2 with positive probability and will continue with d_2, because in the bottom version, $R_2 d_2$ is dominant. Player 2, then, given the move, assesses probability one (at the L_1 action equilibrium) that the bottom version is being played. As 1 will therefore continue with d_2 given the chance, 2 chooses D.

With this example to provide motivation, we turn to a formal development. First, we must modify the definition of an allowable elaboration. The elaboration is said to be an *independent elaboration* if:

(i) for each player i there is an integer N_i such that the number of versions in the elaboration is $\prod_i N_i$;

(ii) writing $(j_1, ..., j_I)$ as the index of one of these elaborations, where $j_i \in \{1, ..., N_i\}$, there is for each i a probability distribution μ_i on $\{1, ..., N_i\}$ such that $\mu((j_1, ..., j_I)) = \prod_i \mu_i(j_i)$;

(iii) there is for each i and $j \in \{1, ..., N_i\}$ a payoff function $u^j(i, z)$ such that i's payoffs in version $(j_1, ..., j_I)$ is given by $u^{j_i}(i, z)$; and

(iv) the information given to i if the true version is $(j_1, ..., j_I)$ concerning which version prevails is simply j_i.

The same convergence criterion as in section 4.2 is used, although we now restrict to independent elaborations for perturbations.

Definition. A strategy σ for an extensive game E is *locally strict under independent elaborations* if there is a sequence of independent elaborations of E, $\{\tilde{E}^k\}$, that converges to E in the sense of the convergence criterion, and a sequence of strict (in the reduced normal form) equilibria $\{\tilde{\sigma}^k\}$ for the respective elaborations, such that the behavior prescribed by the $\tilde{\sigma}^k$ at nodes $(t, 1)$ converges to behavior at node t prescribed by σ.

To give our result, we need a final definition.

Definition. A (pure strategy) equilibrium σ of a normal form game γ is *strictly perfect in the normal form* if there is a sequence of totally mixed strategies $\sigma^k \to \sigma$ such that,

for each player i and index k, σ_i is a *strict* best response to the other players' strategies prescribed by σ^k.

Remark. The reader can easily show that (Ld, D) is strictly perfect in the normal form of the game in figure 5.1.

Proposition 4. For a given extensive form E, every equilibrium σ that is strictly perfect in the normal form is locally strict under independent elaborations.

Proof. Fix a sequence of totally mixed strategies $\sigma^k \to \sigma$ such that each σ_i is a strict best response to each σ^k. In the kth elaboration, N_i is set equal to $\#S_i$, the number of pure strategies of player i. Payoffs for i in versions corresponding to s_i are as in the original game if $s_i = \sigma_i$, and they make s_i dominant otherwise. The marginal probability that type s_i for i is chosen is exactly that assigned to s_i in σ_i^k. Hence these independent elaborations do converge in the sense of the convergence criterion to the originally specified game.

By construction, it is a strict Nash equilibrium in $\tilde{E}^k$ for each player i, if sane, to play σ_i, and to play s_i when crazy of type s_i. This gives the result.

Q.E.D.

References

R. Aumann (1975), "Subjectivity and Correlation in Randomized Strategies," *Journal of Mathematical Economics*, Vol. 1, 67-96.

D. Fudenberg and E. Maskin (1986), "Folk Theorems for Repeated Games with Discounting or with Incomplete Information," *Econometrica*, in press.

E. Kohlberg and J. F. Mertens (1985), "On the Strategic Stability of Games," mimeo, Harvard, forthcoming in *Econometrica*.

D. Kreps, P. Milgrom, J. Roberts and R. Wilson (1982), "Rational Cooperation in the Finitely-repeated Prisoners' Dilemma," *Journal of Economic Theory*, Vol. 27, 245-52.

D. Kreps and R. Wilson (1982), "Sequential Equilibria," *Econometrica*, Vol. 50, 863-894.

R. Myerson (1978), "Refinements of the Nash Equilibrium Concept," *International Journal of Game Theory*, Vol. 7, 73-80.

R. Selten (1965), "Spieltheoretische Behandlung eines Oligopmodells mit Nachfragetragheit," *Zeitschrift fur die Gesamte Staatswissenschaft*, Vol. 121, 301-324.

——— (1975), "Re-examination of the Perfectness Concept for Equilibrium Points in Extensive Games," *International Journal of Game Theory*, Vol. 4, 25-55.

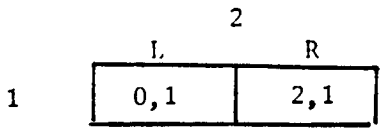

Figure 2.1.

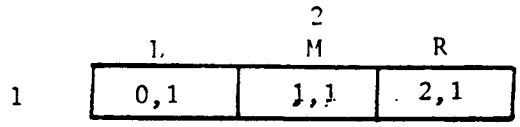

Figure 2.2.

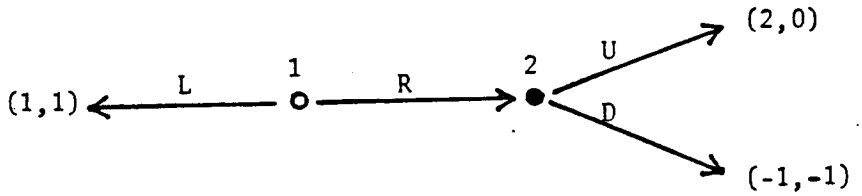

Figure 3.1.

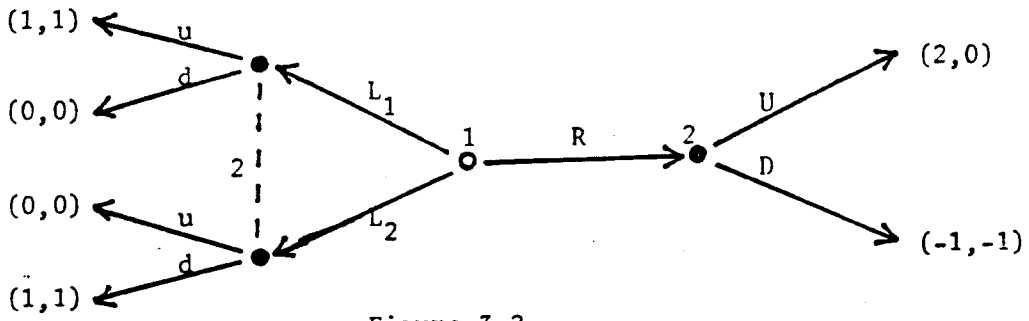

Figure 3.2.

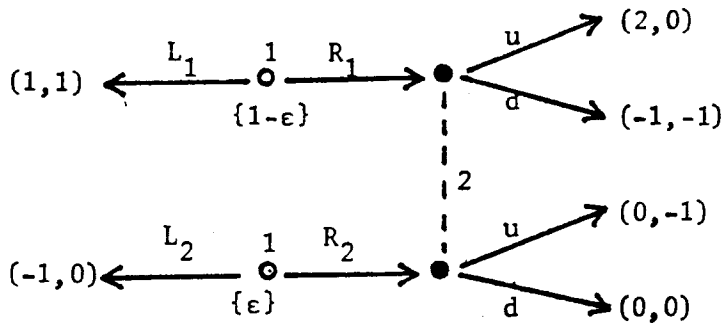

Figure 4.1.

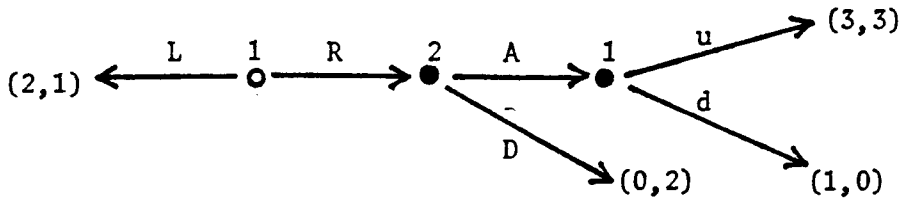

Figure 5.1.

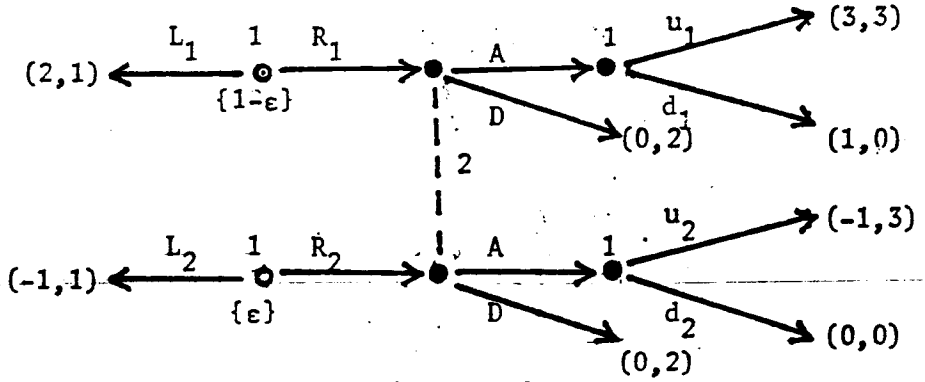

Figure 5.2.